06｜如何将AI技术运用到降噪中？

2021-12-03 冯建元

《搞定音频技术》 课程介绍 >

讲述：冯建元
时长 17:09　大小 15.71M

你好，我是建元。

上节课我们讲了噪声的分类和一些常见的传统降噪算法。传统算法通过统计的方法对噪声进行估计，并可以对稳态噪声起到比较好的降噪作用，但是**在非稳态噪声和瞬态噪声等噪声类型下，传统降噪算法往往不能起到比较好的效果。**

最近几年，随着 AI 技术的不断演进，在降噪等音频处理领域，都出现了很多基于 Artificail Intelligence（AI）或者说基于人工神经网络模型的降噪算法。这些 AI 算法在降噪能力上较传统算法都有很大的提升。但 **AI 降噪算法和很多其它AI 算法一样，在部署的时候也受到诸如设备算力、存储体积等条件的限制。**

这节课就让我们看看 AI 降噪算法是如何一步步实现的，以及在实时音频互动场景中，我们如何解决 AI 降噪算法的部署难题。

AI 降噪模型的基础知识

AI 模型也就是我们经常听到的深度学习模型、机器学习模型或人工神经网络模型。其实 AI 模型的定义更为广泛，后面的这几种说法都是从不同角度描述了目前常用 AI 模型的特点。

AI 模型的构建普遍采用**大量数据训练**的方式，来让模型学习到数据内隐含的信息，这就是所谓的机器学习。**在降噪这个领域，模型的输入是带噪的语音信号，模型的输出是纯净的语音信号**，我们通过大量的这样成对的带噪和纯净的语音数据，来训练 AI 模型，使其具有降噪的能力。

下面我们来看看常见的 AI 降噪模型的结构，以及 AI 降噪模型的训练方法。

常见模型结构

AI 模型常采用人工神经网络来模拟人脑神经的记忆和处理信号的能力。常见的人工神经网络类型有深度神经网络（Deep Neural Network，DNN）、卷积神经网络（Convolutional Neural Network，CNN）、循环神经网络（Recurrent Neural Network，RNN）等。

DNN

一个典型的 DNN 网络结构如图 1 所示：

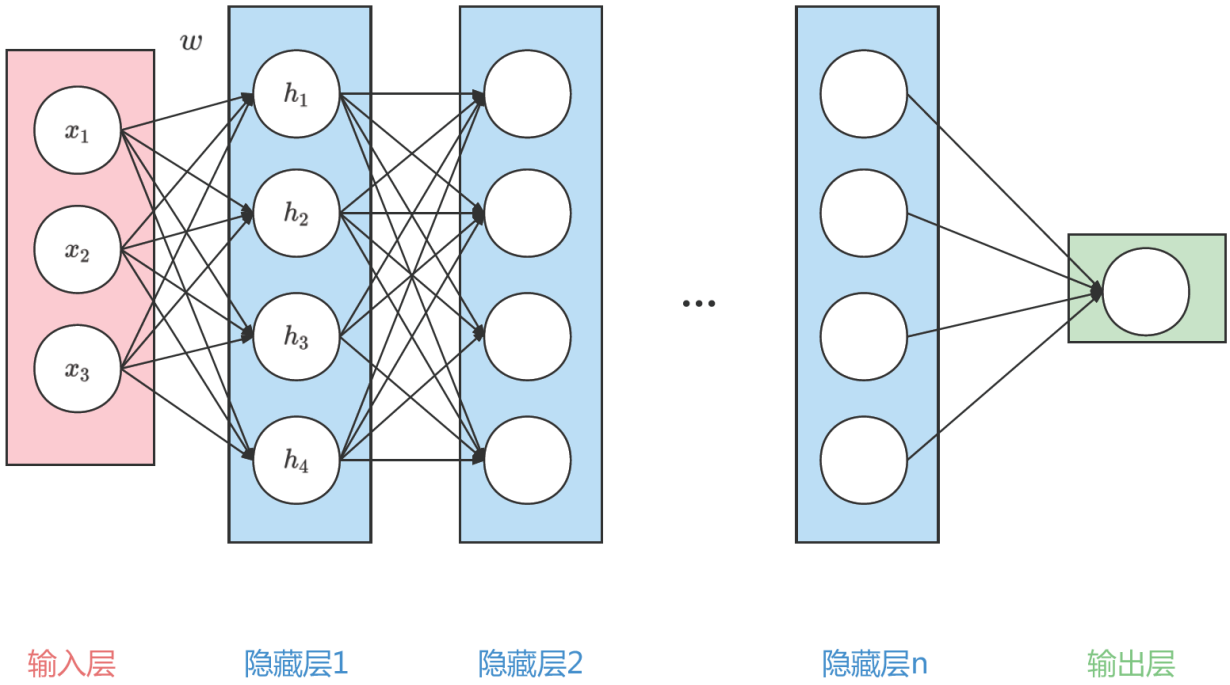

图 1 DNN的结构示意图

可以看到图 1 中信号从输入层到输出层中间经历了 n 个隐藏层。每层都是线性连接，并且每层中的圆圈都代表一个神经元。举个例子，图 1 中隐藏层 1 中的第一个数 h_1，就是由输入层的（x_1，x_2，x_3）的线性加权得到的，即

$$h_1 = w(1,1)x_1 + w(2,1)x_2 + w(3,1)x_3,$$

其中 w 就是第一个隐藏层的权重。在 DNN 的计算中，每个神经元都是前一层的加权平均。这样就可以通过一个多层的线性的网络，来对复杂的信号处理过程建模。

CNN

比较典型的 CNN 网络结构图如图 2.1 和 2.2 所示：

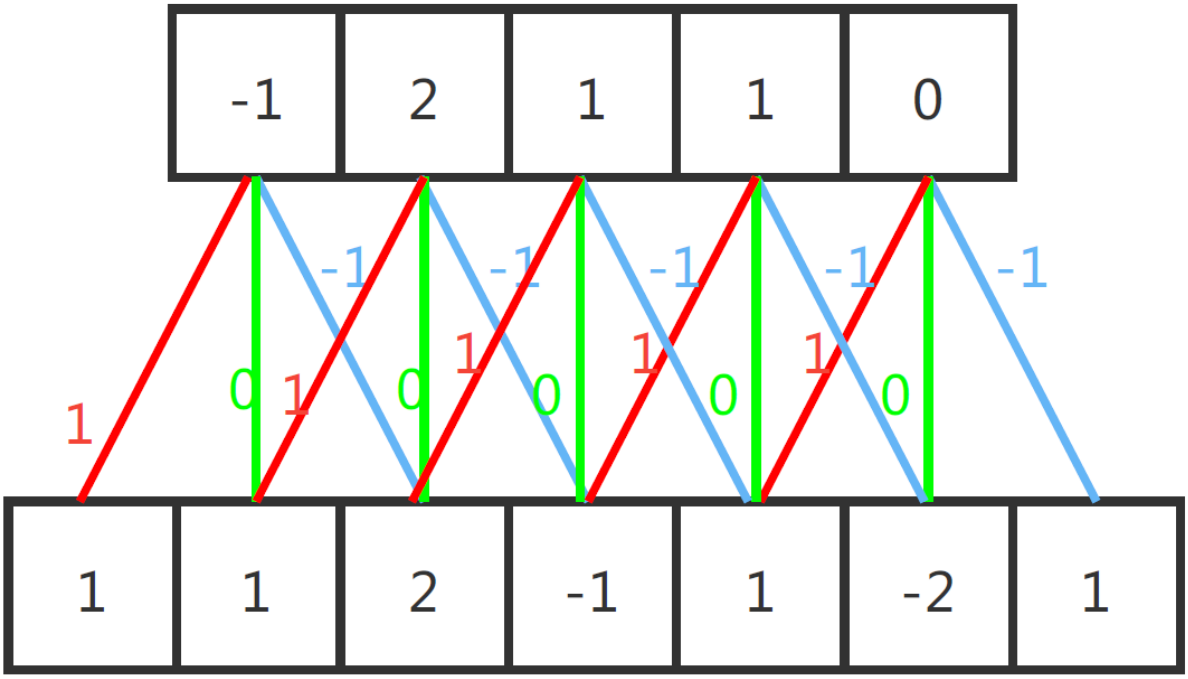

图2.1 一维卷积结构示意图

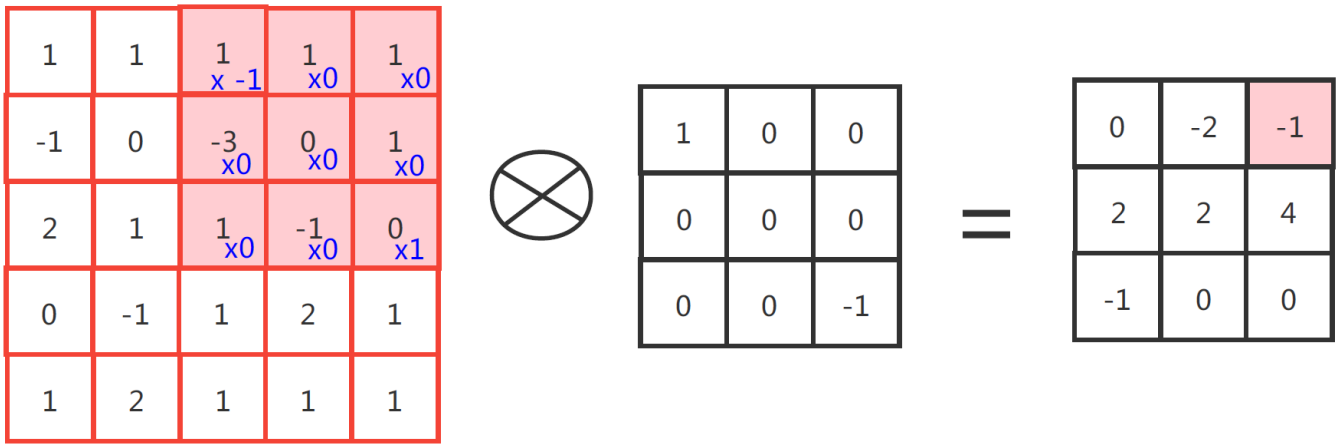

图2.2 二维卷积结构示意图（邱锡鹏 著，神经网络与深度学习）

图 2.1 是 CNN 中的一维卷积的示意图，这里红黄绿三线代表卷积核为（-1,0,1）的卷积计算过程。每一层输出信号都是输入信号和卷积核卷积的结果。比如，输出层中第一个数为

$$1 \times 1 + 1 \times 0 + 2 \times -1 = -1.$$

图 2.2 的二维卷积也是同样的道理，只不过二维卷积中输入、输出和卷积核都是二维的。比如结果里右上角的 -1，就是由标红的输入矩阵与卷积核做点乘，然后再把得到的结果做

累加得到的。

CNN 网络就是由多个这样的一维或者二维的卷积层串联得到的。一维的 CNN 网络，可以直接在一维音频信号上使用，而二维的 CNN 网络最早是用于图像这种二维信号的处理。但其实对音频做了 STFT 后，得到的频谱图也是二维的。所以在频域上做计算时，可以使用二维的 CNN 网络结构。

RNN

典型的 RNN 网络结构如图 3 所示：

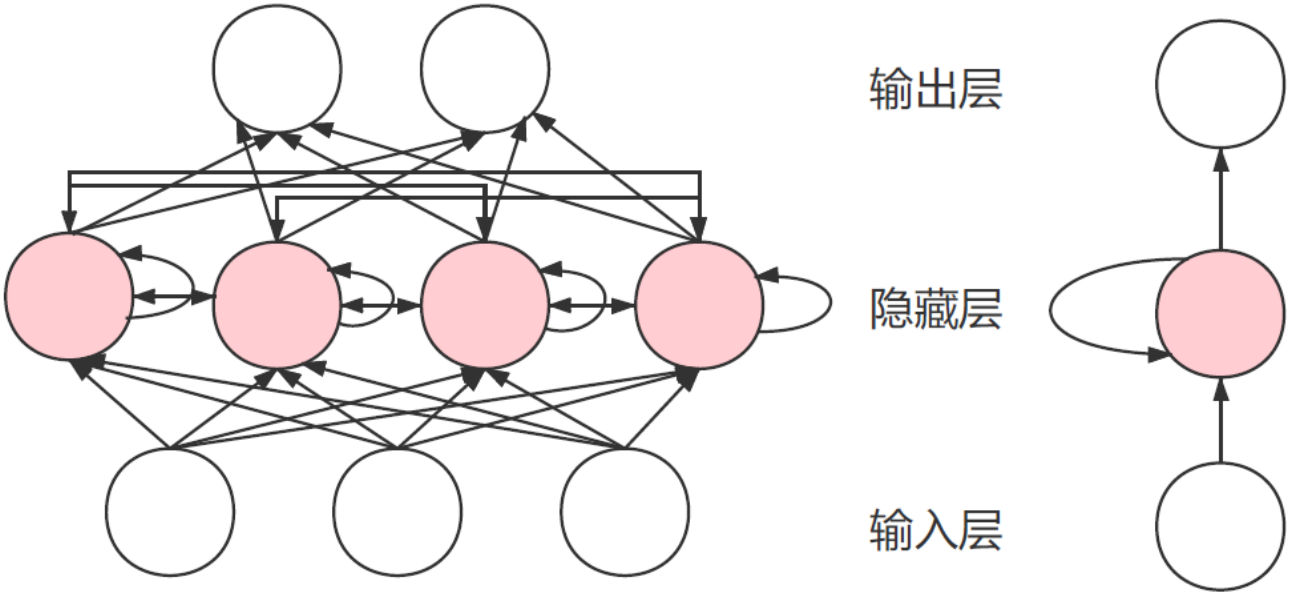

图3 RNN网络结构示意图

可以看到 RNN 网络中隐藏层的每个神经元（粉色圆圈），除了和输入层的信号相关，还和隐藏层本身的状态相关。这种**自回归**的结构是 RNN 的特点。常见的的 RNN 网络有 LSTM，GRU等，由于篇幅所限，这里我们不再介绍。如果你有兴趣可以自行查看一下文献。

从物理含义来解释，如果我们把输入从左到右按照时间来排列，那么 RNN 的自回归特性可以感知信号在时间轴上的特征。换句话说，每个时间点的隐藏特征，除了由自身信号提取，还可以从前后时间点上的信息来得到。所以 RNN 在时序建模中是常见的方法。

模型训练方法

好的，知道了 AI 模型的基本结构，我们来看看 AI 降噪模型是如何训练的。AI 模型中大量的参数，比如 DNN、RNN 中每个神经元的计算权值，以及 CNN 中的卷积核，都需要依靠训练来得到。

所谓训练就是，假设我们给予模型 $y = f(x)$ 一个输入 x，比如 1，然后模型可能会计算出 $\hat{y} = 1.5$，接着我们再告诉模型输出应该是 2。这时模型的误差为 0.5，而模型就会朝着输入为 1 时结果为 2 的方向调整模型的参数值。这样经过多次训练模型就可以拟合出 y 和 x 的之间的映射关系。所以我们只需要准备一组 x、y 作为输入和标签数据，就可以开始训练模型了。

其实 AI 模型训练按照是否有标签数据分类，可分为**有监督的训练和无监督的训练**。而降噪算法万变不离其宗，目的都是将目标信号与噪声信号分离开来。这节课我们讲的 **AI 降噪主要是消除人声之外的所有其它声音**。所以目标信号主要是语音信号。因此，在降噪模型的训练时，**我们一般用的是有监督的训练方式。**

在训练数据里我们一般用纯净的语音作为目标或者说标签，然后用纯净语音加入一些噪声生成含噪数据，作为模型的输入。这里的噪声主要是指环境噪声。回想一下上节课的内容，环境噪声一般为加性噪声，所以在准备训练数据时，我们需要先准备一个纯净语音库和一个噪声库，而含噪的数据可以直接把纯净语音和噪声信号相加来得到。**AI 降噪模型训练**的步骤如下：

1. 通过预处理把含噪数据转换为 AI 模型的输入信号；
2. 通过 AI 降噪模型得到估计的纯净语音信号；
3. 计算模型估计和实际纯净语音信号的差距，也就是常说的 Loss；
4. Loss 通过反向传播，结合梯度下降的方法更新模型的参数；
5. 重复步骤 1～4 直至收敛（也就是 Loss 下降至一个稳定的数值）。

其中，步骤 3 里所说的 Loss，可以用均方差（Mean Suqared Error ，MSE) 等形式。MSE 如下所示：

$$\mathrm{MSE} = \sum_{i=0}^{N} \frac{(s_i - \hat{s}_i)^2}{N},$$

其中 s_i 和 $\hat{s}_i$ 分别代表纯净语音信号和模型估计的语音信号，N 表示信号的长度，模型训练的目标就是最小化模型预估和纯净语音信号的差距。**不同 Loss 的设计会对 AI 模型的结果产生影响**，而在 AI 降噪中，还有很多不同 Loss 的尝试。有兴趣，可以到[参考文献 5]中详细了解。

这样，我们就通过迭代的方法，不断的训练模型，从而得到一组最佳的模型参数。在实际使用的时候，我们就可以用训练好的模型来进行降噪了。那么有了这些基础知识，让我们总结一下作为**一个 AI 降噪模型的设计者要设计哪些东西**：

1. 一个 AI 降噪的模型，包括模型的预处理和后处理流程；
2. 一个合适的 Loss，用于迭代计算模型的参数；
3. 一个合适的语音信号和噪声信号的数据库，用于模型训练。

其实 AI 降噪模型经过这些年的发展，人们已经总结出了一系列比较成熟的方法。基于时域的 AI 降噪算法，输入和输出都是时域的音频信号，无需任何预处理和后处理，可以实现我们常说的“端到端”处理。具有代表性的模型结构有基于 RNN 或 CNN 的 TasNet [参考文献 1, 2]等。

而更多的是基于频域信号，进行建模处理的模型。这类模型是对傅里叶变换后的频域信号进行处理，需要先把原始信号经过 STFT 转换为频谱，然后通过模型和含噪频谱估计出一个纯净语音的频谱，最后需要通过逆 STFT 作为后处理，将频谱转换为时域的音频信号。其中的代表有基于 RNN 的 RNNoise [参考文献 3]，或者结合 CNN 和 RNN 的 CRN 模型[参考文献 4]。如果你有兴趣，可以看看附录里的文献。

了解了 AI 降噪模型的基础知识，接下来我们主要介绍一种最为常用且效果比较好的方法：基于频域掩码的 AI 降噪算法。

基于频域掩码的 AI 降噪算法

在传统降噪中，我们讲的维纳滤波等方法，都是通过计算先验信噪比，然后在频域上对每一个频谱的频点都乘以一个小于等于 1 的系数来抑制噪声。**这些在频域上乘的系数我们统称为频域掩码**。而如何计算这个频域掩码就成了解决降噪问题的关键，传统降噪是基于统计的方法来得到这个频域掩码的，而 AI 算法则是通过人工神经网络模型来对这个频域掩码进行建模的。

基于频域掩码的 AI 降噪算法的主要步骤如下：

1. 带噪的音频信号经过 STFT 得到频域信号；
2. 频域信号作为输入，利用人工神经网络得到频域掩码；
3. 将第 1 步中的频域信号乘以频域掩码，得到降噪后的频域信号；
4. 将降噪后的频域信号做 STFT 的逆变换得到纯净的语音信号。

值得一提的是通过 STFT 后得到的频域信号实际上是复数域的。对复数域的频谱取模就是我们所说的幅度谱（Magnitude Spectrum），它代表不同频点的能量分布。而对复数谱中的实部和虚部的比值求反正切（arctan），可以得到一个取值从 -π到 +π的相位谱（Phase Spectrum）。如果在频谱上乘以一个 0 ~ 1 的实数频域掩码，则修改的就是幅度谱，而相位谱或者说实部、虚部的比值并没有变化。

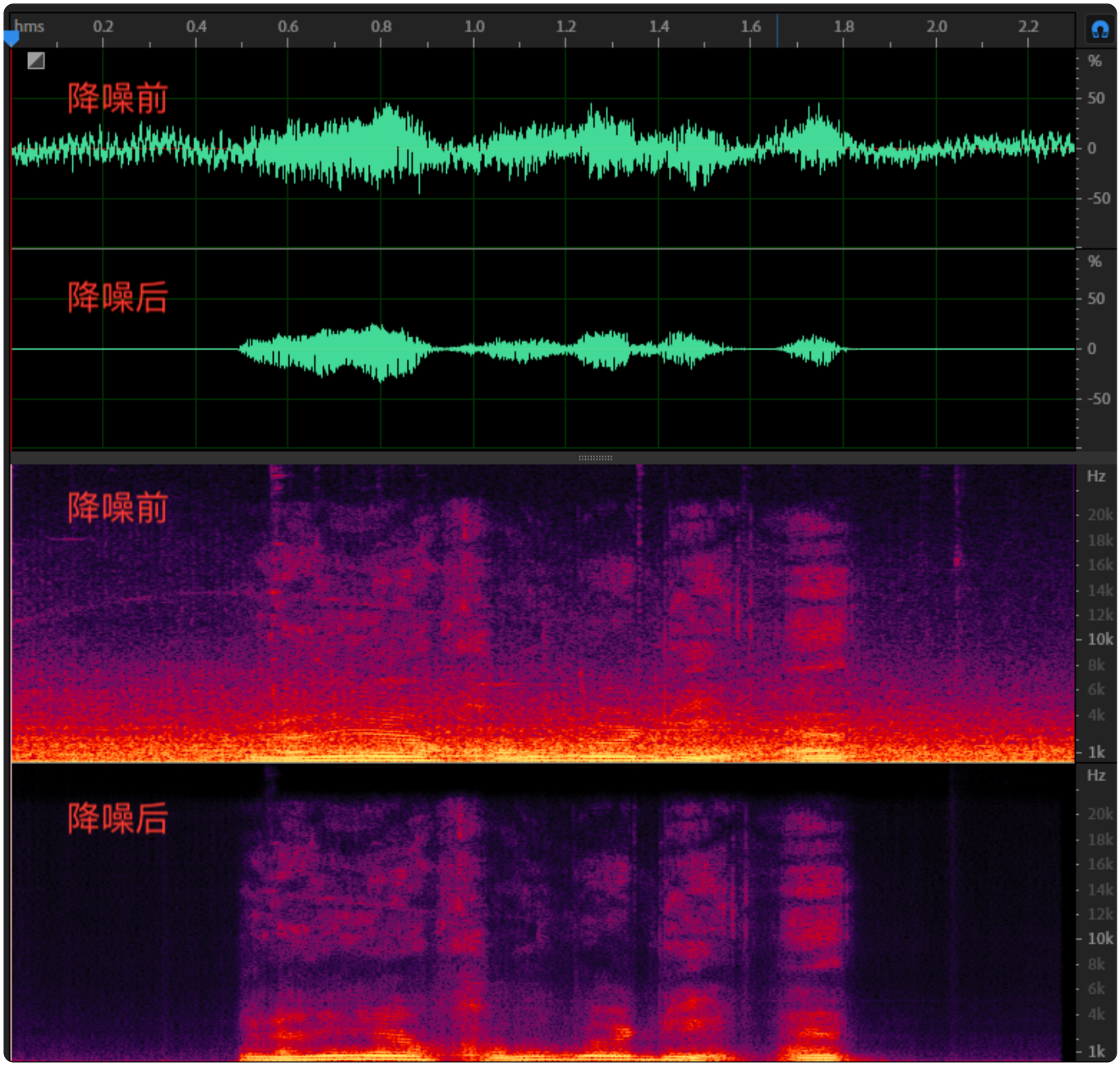

图4 基于频域掩码的AI降噪时域和频域对比图

如图 4 所示就是一个基于频域掩码的 AI 降噪后的对比图，我们可以看到在频谱上噪声的部分能量被抑制了，且在降噪后能看到一个比较清晰的语谱能量分布。

STFT 中相位谱没有可准确描述的物理含义，所以对相位谱的建模会比较困难，而人耳对相位不是很敏感。因此，在传统算法和大部分基于频域掩码的 AI 算法中，都只对幅度谱进行处理，且模型得到的纯净语音和带噪语音的相位还是一样的。虽然人们对相位的差异感知不是很明显，但只改变相位谱的频域掩码不能做到对纯净语音的完美重建。听感上还是能听出一些不同。

近些年，人们开始用 AI 模型来对相位谱或者说对整个复数域的频谱整体进行建模降噪。其中具有代表性的，如微软的 PHASEN 和 2020DNS 降噪比赛中夺冠的 DCCRN 模型等。

若你有兴趣，可以到[参考文献6，7]中自行了解一下。但是在实践中，增加相位谱的恢复相比只对幅度谱做修正，需要消耗更多的算力和模型存储空间，这可能会为模型的部署造成困难。

AI 降噪模型的工程部署

通常 AI 模型在算力和模型参数存储上，都比传统的方法要求要更高一些。现在就让我们看看在实时音频系统中部署一个 AI 降噪模型都需要注意些什么吧。

因果性

在 RTC 等实时音频的应用场景中，降噪处理需要考虑到因果性。也就是说，音频未来的信息是拿不到的。在 AI 降噪模型的选择中，一些双向的网络结构，比如双向的 RNN 模型就不能使用。但语音信号是有短时相关性的，如果一点未来的信息都不用，可能会导致模型的降噪能力下降。

我们可以采用引入一点延迟的方式来提升模型的降噪能力。比如在第 i+m 帧，输出第 i 帧的降噪信号，这样就引入了 m 帧长度的延迟，一般 m 不超过 3。AI 模型的输入可以往前看 3 帧，这种方法也就是我们常说的“look ahead”。

AI 降噪模型存储空间和算力限制

在模型部署的时候，尤其是手机、IOT 等移动端的部署，设备的算力和存储空间都会受到限制。这个需要我们在设计模型结构的时候就加以考虑。模型结构、算力复杂度（Computation Complexity）和参数量（Number of Parameters）之间的关系可参考图 5：

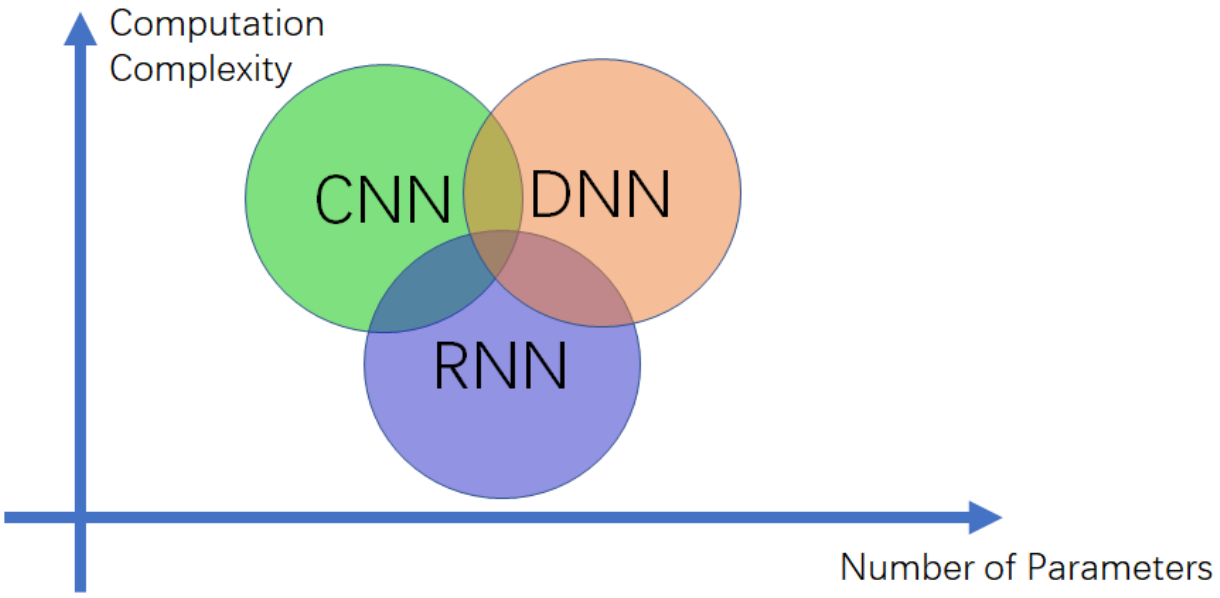

图5 不同模型结构的算力复杂度和参数量分布

在图 5 中我们可以看到 CNN 的参数量最小，这是因为 **CNN 的卷积核是可以复用计算的**。一般基于纯卷积的模型，它的参数量会比较小，而 RNN 和 DNN 本质上都是线性层的计算，所以参数量会比较大。因此，在为移动端等存储空间小的设备设计算法时，会尽量选择 CNN，或者 CNN 结合其它结构的形式来压缩参数量。

另一方面，我们也可以**通过参数量化的方式来对模型进行压缩**。比如，采用 int 8bit 的形式对本来 float 32bit 格式的参数进行量化。注意，参数量化会对模型的精度产生损伤。对于卷积这种可复用的模型结构就不适合做量化，而 RNN、DNN 等结构做量化时对精度的损失就没有那么敏感。

在算力限制方面，我们可以从模型的输入特征着手。比如采用比较小的模型输入，如在 RNNoise 中就是采用 BFCC 这种压缩后的频谱特征作为输入，这样输入信号小了，计算量也就降下来了。另外，刚才说的量化对计算速度也会产生影响。在计算芯片支持不同精度的计算的情况下，量化后的计算速度会更快。模型计算时还可以通过对模型参数和输入数据，按照内存连续读取最优的方式进行重排，来进行计算加速。

我们上面说的量化、加速计算等过程除了自己一个个去完善外，我们在工程部署模型的时候也可以使用一些现成的工具，能帮助我们加速 AI 模型的部署。下表罗列了一些可以使用的 AI 模型部署工具：

工具名称	研发公司
Core ML	Apple
SNPE	QCOM
TensorFlow Lite	Google
Open VINO	Intel

表1 常用AI模型部署工具

小结

好的，让我们来总结一下这节课的内容。AI 模型常用的结构包括 DNN、CNN 和 RNN 等。AI 降噪模型在结构设计时，可以选择其中一种，也可以把这些结构组合使用。AI 降噪模型一般采用有监督的训练方式，并以带噪语音作为模型的输入、纯净语音作为训练的目标。利用反向传播结合梯度下降的方法不断提升模型预估和纯净语音的相似程度。这个相似程度我们一般用，例如 MSE 等形式的 Loss 来表示，并且 Loss 越小，模型得到的结果就越接近于纯净语音。

和传统降噪类似，基于频域掩码的 AI 降噪模型是目前最为常用的 AI 降噪设计。纯净语音频谱的获得，需要对幅度谱和相位谱都进行修正。但如果是在移动端部署 AI 降噪模型，受算力影响，基于幅度谱的 AI 降噪模型可能是最好的选择。

在实时音频信号系统中，降噪模型需要考虑到模型的因果性。在移动端部署时，由于算力和存储空间受到限制，我们需要通过对模型的输入进行降维、模型参数进行量化等操作来进行设备适配。当然我们也可以通过一些现成的工具来快速实现 AI 降噪模型的部署。

在实践中，如果你要自己训练一个 AI 降噪模型，那么数据库（语音、噪声）是不可少的。正好在最近的DNS challenge的降噪比赛里，主办方为我们整理了不少语音、噪声等数据库，有兴趣可自行了解一下。

思考题

其实这里 AI 降噪的模型就是从音频中提取人声部分，如果除了人声之外还想把音乐也保留，那么我们应该怎么设计 AI 模型的输入和输出呢？

你可以把你的答案和感受写下来，分享到留言区，与我一起讨论。我们下节课再见。

参考文献

1、Luo Y, Mesgarani N. Tasnet: time-domain audio separation network for real-time, single-channel speech separation[C]//2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2018: 696-700.

2、Luo Y, Mesgarani N. Conv-tasnet: Surpassing ideal time–frequency magnitude maskingfor speech separation[J]. IEEE/ACM transactions on audio, speech, and language processing, 2019, 27(8): 1256-1266.

3、Valin J M. A hybrid DSP/deep learning approach to real-time full-band speech enhancement[C]//2018 IEEE 20th international workshop on multimedia signal processing (MMSP). IEEE, 2018: 1-5.

4、Strake M, Defraene B, Fluyt K, et al. Fully convolutional recurrent networks for speech enhancement[C]//ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2020: 6674-6678.

5、Ma C, Li D, Jia X. Optimal scale-invariant signal-to-noise ratio and curriculum learning for monaural multi-speaker speech separation in noisy environment[C]//2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC). IEEE, 2020: 711-715.

6、Yin D, Luo C, Xiong Z, et al. PHASEN: A phase-and-harmonics-aware speech enhancement network[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(05): 9458-9465.

7、Hu Y, Liu Y, Lv S, et al. DCCRN: Deep complex convolution recurrent network for phase-aware speech enhancement[J]. arXiv preprint arXiv:2008.00264, 2020.

分享给需要的人，Ta订阅后你可得 20 元现金奖励

生成海报并分享

赞 0 提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 05｜音频降噪如何对症下药？

精选留言 (4)

写留言

Geek_7de4c5
2021-12-03

基于频域掩码的 AI 降噪算法能给个demo吗？

展开

作者回复: https://github.com/xiph/rnnoise，可以看看这个经典的RNNnoise

晓龙
2021-12-03

想要什么声音，就可以将目标数据设置成什么声音，例如需要过滤人声，保留音乐声，就将人声+音乐声作为输入，音乐声作为输出，然后经过AI模型，对比目标，求loss，调整AI模型参数，循环往复，直到梯度下降到一个稳定的数值，这个模型就能过滤人声，保留音乐声。

...

展开

作者回复: 回答的不错

徐刚
2021-12-03

有AI降噪的工程实践例子吗（从训练到工程落地）方便我们小白更加明白ai降噪的操作流程

作者回复: 可以用开源的RNNnoise，练个手https://github.com/xiph/rnnoise

罗 乾 林
2021-12-03

人声+噪音 -> 人声
人声+音乐+噪声 -> 人声+音乐
音乐+噪声-> 音乐
展开

作者回复: 大体如是，其实这里的难点是区分音乐和噪声。这个在行业内也是一个难点